新时代人民日报分词语料库构建、性能及应用(三)
——句长与词的分析比较

黄水清[1,2] 王东波[1,2]
[1] 南京农业大学信息科学技术学院 南京 210095 [2] 南京农业大学领域知识关联研究中心 南京 210095

摘要：［目的/意义］基于新时代人民日报分词语料库从不同维度统计分析句子长度和词汇分布，有助于了解当代汉语文本的语言学特征，进而开展自然语言处理和文本挖掘研究。［方法/过程］在2018年1月人民日报分词语料的基础上，结合1998年1月人民日报分词语料，确定统计中所使用的6种句子类别，统计和分析字与词单位上的句子长度分布，并基于齐普夫定律揭示词汇静态分布情况。［结果/结论］从字词维度上的句子长度分布情况和词汇的齐普夫分布状态上看，随着时间的推移，在1998和2018两个语料上，句子的长度和词汇的分布均发生变化，但这种变化又是延续的、有关联的。

关键词：新时代人民日报分词语料 语料库 句子长度 词汇分布 齐普夫定律
分类号：G255.1
DOI:10.13266/j.issn.0252-3116.2019.24.001

1 引言

1.1 研究背景和意义

从语言学和统计学的角度对文本构成的句子和词汇进行统计和分析是开展语言分析和文本处理的前提。句长的厘定对语言学研究来说有助于语言规律的挖掘、二语习得教学的展开，对机器学习来说有利于特征的统计和选取。词汇分布情况的统计不仅有益于人们了解自我使用语言词汇的概貌，而且也为汉语的自动分词、词性标注和句法分析提供了最为直接的词汇特征知识。

鉴于此，基于自行构建的新时代人民日报语料(New Era People's Daily Segmented Corpus，简称NEPD)，结合北京大学1998年1月人民日报语料，对前后跨度20年的人民日报语料进行字和词维度上的句子长度分布和词汇分布的统计与分析。这一探究不仅从句子长度和词汇分布上全面展示前后两个语料的相同之处，而且也呈现了其不同之点。NEPD涵盖了《人民日报》2015上半年(1-6月)及2016年1月、2017年1月、2018年1月共9个月的语料，同时进行了人工分词标注，是经过人工加工的精语料[1-2]，以学术研究为目的的人员可访问网站http://corpus.njau.edu.cn提交申请免费获得。本文所使用的是其中的2018年1月语料。

1.2 相关研究情况梳理

句长作为重要的语言特征，具有较高的研究价值。无论是语言的纵向发展还是不同语言间的横向比较，从字和词的角度分析句长都是其中关键的研究要素。关于句长的研究，目前主要集中在非特定领域的句长规律分析和语言教学特定领域的句长探究这两个方面。关于非特定领域句长规律的代表性研究如下：基于对赫西俄德和荷马等现存著作中句长进行分析，D. Clayman[3]证明了句子长度和音素数量的分布与作品风格和主题存在着极大的联系。这一研究成果对作者所属艺术流派的鉴别判断有重要的借鉴作用。通过对120万字汉语语料中所有句子进行切分和统计，黄自然[4]分别从“字”和“词”的维度统计了句子长度的分布，总结了句长和平均句长频次分布的内在规律。在结构和使用细节上，古汉语和现代汉语存在着较大差异，李青苗[5]基于对《左传》中的偏正结构和句子长度的统计分析，从句法分析的角度证明了汉民族对细节意义逐步重视，从而使得现代汉语比先秦古代汉语在句长上增加非常多。在句长研究的基础上，王萍与石锋[6]结合“阅读语调”对汉语普通话中不同语句类型的阅读时长表现进行了细致地分析和比较，并归纳出了“阅读时长”的分布模式，该研究是对句长研究的拓

作者简介：黄水清(ORCID:0000-0002-1646-9300)，教授，博士生导师，E-mail:sqhuang@njau.edu.cn；王东波(ORCID:0000-0002-9894-9550)，教授，博士生导师。
收稿日期:2019-11-20 本文起止页码:5-15 本文责任编辑:王传清

展和补充。在张绍麒和李明[8]对汉语句长的研究基础上,左思民[7]指出人脑的信息加工处理方式、语调以及某些强制性的表达格式是制约句长增长的三大因素。

另一方面的相关研究主要集中在语言教学上,尤其是儿童语言的教学上。句长在评价儿童句法发展时是一个关键性的指标,在儿童母语习得研究、智障儿童语言习得研究和第二语言习得研究上均有重要应用。从词和语素的角度,黄自然和贾成南[9]统计了不同年龄段儿童的平均句长,并在实验中将平均句长与句法结构复杂度结合起来,证明了平均句长作为评测儿童语言发展的有效性。与上述研究类似,在针对儿童词汇广度和句法复杂度的实验研究中,金志娟和金星明[10]与 M. Montgomery 等[11]均分析得出了学龄前儿童的语言整体发展状况。学界对句子长度的研究不仅局限于单一语言的历史发展情况和儿童语言的习得,而且在第二语言的学习过程中句长的研究也具有较强的价值和意义。通过对中美高中生的英文写作结果进行统计分析,李建平和张晓蕾[12]发现中国中学生写作的平均句长低于美国中学生写作的平均句长,且不同长度句子在分布上存在很大差异。该研究从二语习得的角度证明了句长研究的现实价值。

在本文的研究中,除了探究句子长度的分布之外,对于词汇自身的分布状况结合齐普夫定律也进行了探究。齐普夫定律是美国学者齐普夫[13]于20世纪40年代提出的词频分布定律,在90年代早期,国内一些学者对该定律进行了一些探索和研究。首先是对于齐普夫定律的探究,比较有代表性的研究如下:冯志伟[14]对齐普夫定律的来龙去脉作了说明,同时指出面临同频词时,同频词的排序等级还有待进一步研究。孙清兰和王肇建[15]对上述问题进行了探讨,并通过理论研究和实例验证证明了最大值法为齐普夫定律中确定词级的最优方法。为了揭示词出现的频次与同频词数量的制约关系,孙清兰[16]提出了新的高频、低频词界分公式,并通过理论分析和实验验证表明了其可靠性。W. Li [17]对齐普夫定律的数学验证方式作了总结和归纳。其次,在齐普夫定律的验证和应用方面,相关学者也做了许多工作。基于由1949年至2008年间意大利十位总统的年终演讲组成的语料,A. Tuzzi 等[18]统计发现即使文本的编撰者不止一人,词汇分布依然遵循齐普夫定律,而且不同人的语言风格也能在词频上清晰地体现出来。沈关龙[19]统计了英文文献《国外电力可靠性文献检索系统》中篇名的标题词频,对齐普夫第一定律和齐普夫第二定律低频区词频分布理论做了验证。基于《科技情报工作的科学技术》语料,王崇德和来玲[20]通过计算C值的中心特征,说明中文文集中的词汇呈现了齐普夫分布规律。通过对中文小说《坚硬的稀粥》的前18个段落进行词频统计与分析,何风远[21]发现词频在分布上呈现出较为明显的齐普夫分布规律。最后,齐普夫定律在生活中普遍存在,除了自然语言领域,一些学者将齐普夫定律的应用延伸到了城市规模[22]、公司规模[23]、网站访问量[24]以及文献计量学等领域,依然可以得到大量符合齐普夫定律的场景。在上述国内外研究的基础上,本文基于2018年1月人民日报分词语料,结合1998年1月人民日报分词语料,从字、词维度的句子长度分布与词汇的齐普夫分布这两个角度,进行系统而全面地统计与分析,并对相应结果进行对比研究。

2 字维度上的句长分布

为了更加充分和全面地统计人民日报语料中句长的分布情况,本研究按照"。(句号)""?(问号)""!(感叹号)"";(分号)"":(冒号)""……(省略号)"这6类标点符号统计人民日报语料中以字和词为单位的句长整体分布情况。对于非上述6类标点符号结尾的句子统一归属到其他类中。其他类主要包括部分特殊表达以及人物对话,如:"(新华社记者丁仁贵摄)""图片新闻""二、原因和教训""上海综合指数周K线图"和"一次公司下班后开会,会议结束前,徐柏玉问:'哪一位夜晚从家外出不关灯?请举手!'"等。其中,以上述人物对话为例,按本研究的统计规则将其分为4句话:①一次公司下班后开会,会议结束前,徐柏玉问:②哪一位夜晚从家外出不关灯?③请举手!④"哪一位夜晚从家外出不关灯?请举手!",由于引号内的两句话已计入问号类别和叹号类别,故将第4句归为其他类,在后文的分析中不予统计。1998年1月和2018年1月人民日报语料中不同类别句子分布情况如表1和表2所示:

表1 1998人民日报语料不同类别句子分布情况

句子类别	句子数量(句)
。	35 982
;	2 636
?	761
!	664
……	346
:	3 258
其他	7 173
总计	50 820

表 2　2018 人民日报语料不同类别句子分布情况

句子类别	句子数量(句)
。	75 450
;	4 867
?	2 071
!	927
……	236
:	4 211
其他	7 370
总计	95 132

2.1　1998 年 1 月人民日报字维度上的句长分布

在《人民日报》1998 年 1 月数据的基础上,以"。""?""!"";"":""……"为标点符号得到 6 类句子,共 43 647 句。以字作为句子的基本构成单位,统计并计算句子的长度,选取句子长度出现频次前 20 名的句子长度分布情况如表 3 所示:

表 3　1998 年句子长度整体分布

序号	句子长度(字)	频次(次)	序号	句子长度(字)	频次(次)
1	26	940	11	22	843
2	29	915	12	17	837
3	23	909	13	32	837
4	21	895	14	35	830
5	18	890	15	20	820
6	34	886	16	31	811
7	25	883	17	30	807
8	24	862	18	36	806
9	28	857	19	19	803
10	33	845	20	27	801

对句子长度进行统计,排名前 20 的句子长度频次均超过了 800 次,且句子长度全部分布于 17-36 字之间。出现频次超过 800 次的各种长度句子频次总和为 17 077 次,占总体的 39.13%,超过三分之一,具有较高的代表性。且长度为 26、29 和 23 字的句子出现频次均超过 900 次,三者共出现 2 764 次,占总体的6.33%,占前 20 名的 16.19%,占总体长度分布的较大比重。但从总体来看,长度为 17-36 字之间的句子出现频次分布较为平均,未出现大幅度波动。

以"。""?""!"";"":""……"为句子分隔符,分别统计 6 类句子的长度分布情况,展示不同句子分割符在语料中出现的频次及占比情况。六类句子长度具体分布情况见表 4。

以句号为分隔符的句子共出现 35 982 次,以句子长度出现频次为排序依据,得到出现频次前 20 名的句子长度分布情况,出现频次总和为 14 158,占全体总数的 39.35%,接近 40%,占比较高。其长度均分布于 20-41 字区间内,且出现频次超过 700 次的句子长度共出现 8 899 次,为总体的 24.73%,占前 20 名的 62.85%,较为明显地分布于 21-36 字区间内。与总体情况相比,以句号为分隔符的句子长度分布情况呈现出较大的起伏,句子长度为 26 字的句子出现频次最多,为 813 次,而长度为 41 字的句子仅出现 618 次,更加证明了句子长度分布的集中趋势,即长度为 21-36 字之间的句子是以句号为分隔符的句子中的重要研究对象。

分号在句子中担任分隔符时与句号作用相似,其长度分布情况也与句号分隔符有较高相似度。以分号为分隔符的句子共出现 2 636 次,远低于以句号为分隔符的句子数量。出现频次前 20 名的句子长度分布于 12-35 字的区间内,共出现 1 236 次,占总体的 46.89%,句子长度集中程度较高,分布的集中趋势较明显,且分布较平均,未出现较大差距。这与分号在句子中的功能有较大联系,分号是介于逗号和句号之间的标点符号,主要用以分隔存在一定关系(并列、转折、承接、因果等,通常以并列关系居多)的两句分句,故分号左右的句子存在一定的结构相似性,长度差距也控制在较小范围内。

感叹号作为特殊的句子分隔符,多出现于表达强烈情感的句子中,总体出现频次较少,共出现 664 次,占总体的比例为 1.52%,远低于以句号和分号为分隔符的句子数量。该类别中,句子长度出现频次排名前 20 的句子,长度分布于 3-30 字区间中,最短句子长度远低于其他类别,且句子长度小于 10 字的句子数量为 154 句,占总体的 23.19%,占前 20 名的 42.42%。感叹号为分隔符的句子多为新闻稿中简短但语气强烈的短句,长度分布趋势较为明显。

冒号通常表示提示语后的停顿或表示提示下文或总括上文,使用较为普遍,以冒号为分隔符的句子共出现 3 258 次,占总体的 7.46%,占比仍远低于以句号和分号为分隔符的句子数量,且句子长度出现频次排名前 20 的句子长度分布于 3-23 字之间,总体来说句子长度较短,且长度不超过 5 字的句子出现总频次为 588 次,占总体的 1.35%,占排名前 20 的长度的句子数量的 25.06%,超过四分之一。

以省略号为分隔符的句子共有 346 句,数量远低于其他类型的句子,其占全部句子数量的 8%,比重较小,仅出现在极少数情况下,表达语义难尽或断断续续等含义。其长度主要分布于 13-43 字区间内,起伏偏差较大,但出现频次较相似,均不高于 10 次。

表 4 1998 年 6 类句子长度具体分布

序号	句号		分号		感叹号		冒号		省略号		问号	
	长度(字)	频次(次)	长度(字)	频次(次)	长度(字)	频次(次)	长度(字)	频次(次)	长度(字)	频次(次)	长度(字)	频次(次)
1	26	813	20	78	11	28	3	280	27	10	11	45
2	34	779	17	70	6	26	18	207	34	10	8	40
3	29	771	22	70	7	26	5	187	16	9	10	37
4	23	755	12	66	4	25	17	173	21	9	13	30
5	24	737	21	66	9	25	11	124	43	9	12	28
6	25	733	26	65	10	21	4	121	22	8	19	28
7	33	732	25	64	3	19	10	113	33	8	14	26
8	32	726	33	64	12	19	12	102	35	8	18	26
9	35	718	16	63	14	19	9	98	13	7	7	24
10	21	712	28	63	16	19	15	94	17	7	15	24
11	28	712	30	63	15	17	8	93	20	7	9	23
12	36	711	24	61	23	15	19	93	25	7	17	21
13	31	695	23	59	13	14	16	92	29	7	20	21
14	22	689	35	59	19	14	14	90	31	7	21	19
15	30	674	14	57	25	14	6	88	38	7	25	18
16	38	665	27	57	28	13	13	87	15	6	6	17
17	27	664	29	55	30	13	20	84	19	6	30	17
18	37	629	32	55	8	12	21	81	24	6	5	14
19	20	625	18	51	18	12	7	77	26	6	16	14
20	41	618	19	50	29	12	23	62	30	6	23	13

问号是语气语调的辅助符号工具,表示一句话结束之后的停顿,用于疑问句、设问句和反问句结尾,新闻报道中常用于反问句表达强烈感情。以问号为分隔符的句子数量为 761 句,占总体的约 1.74%,总的来看比重较低。出现数量排在前 20 位的句子长度分布于 5 –30 字,共出现 485 次,占全部以问号结尾的句子的 63.73%,比重较高,超过以上 5 种类型句子的前 20 名频次所占比重,句子长度的集中趋势最明显,故该长度区间的句子有重要的研究价值。

2.2 2018 年 1 月人民日报字维度上的句长分布

在 2018 年 1 月人民日报语料的基础上,同样以"。""?""!"";"":""……"6 种标点符号为句子分隔符,本文共获得了 87 762 个句子。统计不同长度句子的出现频次,对频次降序排列,并取出现频次前 20 名的句子长度为示例和研究对象,如表 5 所示:

表 5 2018 年句子长度分布

序号	长度(字)	频次(次)	序号	长度(字)	频次(次)
1	26	1 668	11	23	1 571
2	27	1 634	12	28	1 548
3	33	1 621	13	31	1 547
4	25	1 602	14	36	1 511
5	34	1 599	15	22	1 498
6	30	1 595	16	20	1 479
7	29	1 595	17	38	1 454
8	35	1 594	18	19	1 448
9	24	1 592	19	37	1 432
10	32	1 578	20	21	1 426

排在前 20 位的句子长度出现频次均超过了 1 400 次,总出现频次为 30 992 次,占全部句子数量的 35.31%,超过总数的三分之一。表 5 中句子长度分布于 19 –38 字的区间中,长度跨度较大,且出现频次落差较大,如长度为 26 字的句子出现 1 668 次,比长度为 21 字的句子出现频次多 242 次。出现频次前 20 名的句子长度的频次平均值为 1 549.6 次,与中位数差距较小,变化曲线较平稳。同时从句长排序居前三的数据分布来看,2018 年 1 月的句子在长度上要比 1998 年 1 月有所增加。

对 6 种不同类型的句子长度分布情况进行统计,并以句子长度出现频次降序排列,选取排在前 20 位的句子长度及出现频次见表 6。

句号作为分隔符的句子出现频次为 75 450 次,占全部句子数量的 85.97%,接近五分之四,比例极高,句子长度出现频次前 20 位的数量总和为 25 311 次,占全部句子总和的 28.84%,同时占以句号为分隔符的句子总数的 33.54%,超过了三分之一,且句子长度集中在 22 –44 字区间内。与 1998 年数据相比,以句号为分隔符的句子比例大幅提升,增长 118%,句子长度区间未出现较大变化,较为明显地展示出人民日报使用句号的频率大幅提升,语言表达方式出现较大变化。

表 6　2018 年 6 类句子长度具体分布

序号	句号		分号		感叹号		冒号		省略号		问号	
	长度(字)	频次(次)	长度(字)	频次(次)	长度(字)	频次(次)	长度(字)	频次(次)	长度(字)	频次(次)	长度(字)	频次(次)
1	26	1 413	24	129	9	37	3	248	58	11	12	103
2	33	1 404	27	119	8	30	5	202	22	11	14	82
3	35	1 398	19	118	10	30	4	199	38	10	13	77
4	34	1 393	28	112	13	30	9	149	23	9	15	72
5	30	1 388	22	111	17	29	6	143	26	8	20	71
6	27	1 380	20	111	19	29	11	128	42	7	19	71
7	32	1 361	29	108	12	26	10	127	51	7	10	65
8	29	1 360	21	107	7	26	7	124	27	7	11	64
9	25	1 356	26	106	22	26	8	120	33	6	18	63
10	36	1 330	25	105	20	24	15	116	28	6	16	63
11	31	1 322	30	104	24	24	13	114	34	5	17	62
12	23	1 316	35	103	18	24	14	110	24	5	21	62
13	28	1 302	32	102	15	21	12	106	8	5	8	59
14	24	1 296	34	102	16	21	18	106	32	5	7	55
15	38	1 294	23	101	23	21	19	106	18	4	23	53
16	40	1 274	31	100	6	20	16	103	31	4	25	51
17	39	1 259	16	95	14	19	20	95	36	4	9	50
18	37	1 253	41	95	35	19	17	95	17	4	24	46
19	22	1 212	33	94	33	17	22	93	47	4	22	45
20	44	1 204	36	94	5	17	24	92	25	4	6	43

以分号为分隔符的句子共出现 4 867 次,占总体数据的 5.54%,句子长度主要分布于 16－41 字区间内,与 1998 年同月数据相比,该类型句子占全部句子数量的比例降低约 9%,变化较小,句子长度区间的上下限均有所提高。总体来说,2018 年数据中以分号为分隔符的句子占比略低于 1998 年的同月水平,句子长度有所提升,以分号结尾的复杂长句使用更加频繁。

以感叹号为分隔符的句子在 2018 年的数据统计中仍占比较小,共出现 927 次,占总体数量的约 1%,是 1998 年同期数据占比的三分之二左右,但由于基数较小,总体差距不大,与 1998 年数据相比,使用感叹号表达强烈语气的句子仍占极小部分,且使用比例有所降低。句子长度区间由 1998 年的 3－30 字提升到 5－35 字,增幅较小,变化不大。

冒号作为句子分隔符获得的句子数量仍是以句号和分号为分隔符的句子数量之外最多的,共4 211 句,占全部句子总数的 4.80%,只有 1998 年同期数据的 64.34%,降幅较大,以冒号开启下文的句子比例下滑。与此同时,句子长度所在区间为 3－24 字,与 1998 年数据基本相同,冒号的使用习惯变化较小,但使用频率大大下降。

从统计数据来看,以省略号结尾的句子数量为 236 句,占全部句子总数的 2%,在总体句子数量基数增加的条件下,以省略号为分隔符的句子数量仍呈下降趋势,与 1998 年同期数据占比相比下降 75%,由 8% 的占比下降到仅占总数的 2%。较为明显地说明人民日报报道中语义难尽和断续说明等表达方式正在被明确详细的表达方式所取代。

以问号作为分隔符的句子数量为 2 071 句,占总体数量的 2.36%,其中排名在前 20 名的句子长度出现频次为 1 257 次,占总体数量的 1.43%,占以问号为分隔符的句子总数的 60.70%,句子长度主要分布于 6－25 字之间。与 1998 年同期数据相比,以问号结尾的句子总数和占比均有小幅提升,前 20 名句子长度占该类型句子总数的比例略降低,句子长度间差距变小。

3　词维度上的句长分布

在上述字统计的基础上,按照已经分词后的结果,本研究统计和分析 1998 年 1 月和 2018 年 1 月人民日报词汇维度句长的整体分布情况。

3.1　1998 年 1 月人民日报词维度上的句长分布

与字基础上的长度分布一致,本文在句子长度分布统计时,去除了类别为其他的文本。统计发现 1998 年 1 月的句子中,句子长度分布状态为在 2－198 词的

区间离散分布。句子长度的分布如图 1 所示，图 1 中的横坐标为句子长度，纵坐标为对应长度的句子出现的频次，即句子数量。

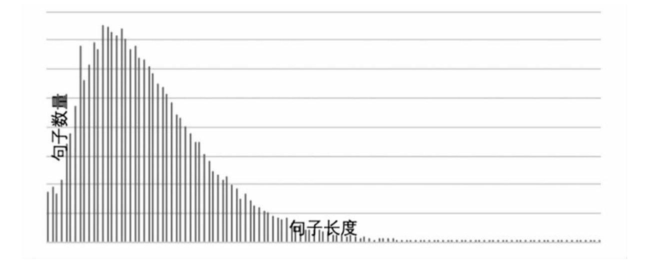


图 1　1998 年人民日报语料句子长度区间分布情况表 – 总(以“词”为单位)

从图 1 中按照句子长度频次即句子数量降序选取频次最高的前 20 个句子长度，具体分布情况如表 7 所示：

表 7　1998 年人民日报语料句子频次前 20 句子长度分布情况 – 总(以“词”为单位)

序号	句子长度(词)	频次(次)	占比(%)
1	14	1 510	2.97
2	15	1 498	2.95
3	18	1 478	2.91
4	16	1 455	2.86
5	17	1 437	2.83
6	19	1 406	2.77
7	12	1 391	2.74
8	21	1 367	2.69
9	9	1 363	2.68
10	20	1 337	2.63
11	13	1 335	2.63
12	22	1 282	2.52
13	23	1 266	2.49
14	11	1 227	2.41
15	24	1 225	2.41
16	25	1 174	2.31
17	10	1 120	2.20
18	26	1 104	2.17
19	27	1 082	2.13
20	28	1 031	2.03

从图 1 中可以看出，1998 年的人民日报语料句子共有 50 820 句，以词为单位分布范围为 2 – 198 词。句子长度的分布呈现明显的集中情况，结合表 3 可以得到，句子长度主要集中在 9 – 30 词这个区间范围，这个句长范围的句子出现频次都在 1 000 次以上。1998 年 1 月人民日报句子的整体分布有着明显的“拖尾”现象，从 30 往后呈现下降的趋势，最后出现大量的 1，即不同长度句子出现频次只有 1 次。前 20 的句子数占了总句子数的 51.33%，前 20 的句子长度的句子数在总句子数中的占比均超过 2%。从频次排 21 的句子长度开始，各个长度的句子的数目占总数目的比重都低于 2%，将前 20 的句子长度按频次来区分可以分为频次大于 1 300 次和频次小于 1 300 次两个区间，第一区间——1 – 11 名占总句子数比重 30.65%，第二区间——12 – 20 名占总句子数比重 20.68%。将前 20 的句子长度按照句子长度所属区间可以分为 1 – 10、11 – 20、21 – 30 三个区间，前 20 中长度在 1 – 10 词的有 9、10，长度在 11 – 20 词的句子均包含在其中，长度在 21 – 30 词的有 20、22、23、24、25、26、27、28。可以看出，句子长度主要集中在 11 – 30 词之间。在整体的长度分布当中，对 1998 年 6 类频次居于前 10 的句子分布情况进行了具体分析，具体分布见表 8。

将句子按照句号、分号、感叹号、冒号、省略号和问号来划分，依据前 10 的句子长度分布来看：以句号结尾的句子长度主要集中在 14 – 23 词；分号结尾的句子长度主要集中在 8 – 19 词；感叹号结尾的句子长度主要集中在 3 – 16 词；冒号结尾的句子长度主要集中在 2 – 14 词；省略号结尾的句子长度主要集中在 10 – 24 词；问号结尾的句子长度主要集中在 5 – 16 词。与整体的前 20 高频句子长度数分布相比较，以句号结尾的句子的高频句长中没有 9、10 这两个长度数，多了 29，其余与整体一致；以分号结尾的句子中出现了 7、8 这两个句子长度数，这与整体的高频句子长度数略有不同，其余也均与整体一致；感叹号的句子中与整体不一致的句子长度数是 2、3、4、5、6、7、8；冒号、问号中也都有一些较小的句子长度数。以除句号外的其他符号为结尾的出现频次前 20 的句子长度数与整体相差较大，但由于以分号、感叹号、冒号、省略号和问号结尾的句子只有很少一部分，因此对最终的分布影响较小。结合图 1，以分号、感叹号、冒号、省略号和问号结尾的句子虽然只有很少一部分，但是其高频句子长度却在句子长度最集中的区间内。

3.2　2018 年 1 月人民日报词维度上的句长分布

按照上述方法和流程，本文基于词这一单位统计 2018 年 1 月人民日报语料中句子的分布情况。句子长度分布状态为 2 – 309 区间的离散分布。图 2 为句子长度的分布图，横坐标为句子长度，纵坐标为对应长度的句子出现的频次，即句子数量。从图 2 中按照句子长度频次即句子数量的降序选举频次最高的前 20 个句子长度，得到的分布情况见表 9。

表8　1998年人民日报语料句子长度区间分布情况－分(以“词”为单位)

序号	句号		分号		感叹号		冒号		省略号		问号	
	句子长度（词）	频次（次）	句子长度（词）	频次（次）	句子长度（词）	频次（次）	句子长度（词）	频次（次）	句子长度（词）	频次（次）	句子长度（词）	频次（次）
1	18	1 272	13	117	9	37	9	439	21	16	6	63
2	16	1 255	17	116	3	35	2	310	14	14	7	56
3	15	1 252	19	110	6	34	3	291	17	13	8	51
4	14	1 219	14	107	7	30	6	243	24	13	9	46
5	17	1 212	15	107	5	29	8	167	10	12	10	39
6	21	1 193	18	103	8	29	4	163	12	12	13	33
7	19	1 192	9	100	4	27	7	159	20	12	15	32
8	20	1 169	10	100	16	26	11	144	15	11	11	29
9	23	1 127	12	100	10	23	5	140	18	11	16	29
10	22	1 124	8	93	11	22	14	140	22	11	5	27
11	12	1 121	24	90	22	22	12	113	30	11	12	26
12	13	1 069	20	87	19	20	10	99	19	10	4	25
13	24	1 055	21	86	12	19	13	98	23	10	17	19
14	25	1 048	11	83	18	19	15	79	29	10	24	19
15	27	989	7	82	15	17	17	67	32	10	14	18
16	26	981	22	75	21	15	16	65	28	9	3	17
17	11	942	16	72	25	15	19	57	2	8	18	17
18	28	929	25	71	26	14	18	56	13	8	19	17
19	29	887	26	70	2	13	21	45	16	8	23	17
20	18	1 272	13	117	9	37	9	439	21	16	6	63

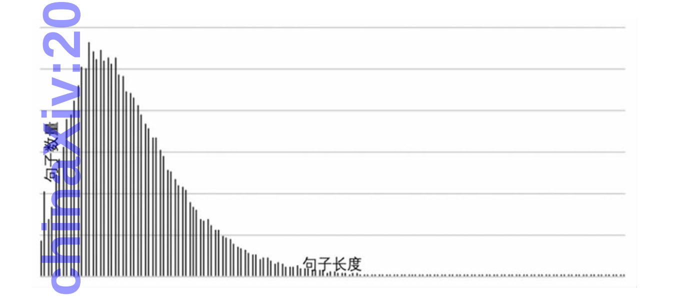


图2　2018年人民日报语料句子长度区间分布情况－总(以“词”为单位)

2018年的人民日报语料句子共有95 132句,以词为单位分布范围为2－309词。与1998年人民日报语料相似,句子长度的分布呈现明显的集中状况,频次最高的是长度为15词的句子,共有2 834句,排名前20长度为29词的句子也有1 961句。结合图2和表9可以看出,句子长度主要集中在10－30词这个区间范围,这个句长范围的句子数都在1 900句以上。相较于1998年人民日报语料,2018年人民日报语料的语料量有很大的增长。

2018年人民日报句子的整体分布也有着明显的“拖尾”现象,句子数在句子长度为15词达到顶峰,随后呈下降趋势。前20的句子数占了总句子数的50.55%,

表9　2018年人民日报语料句子频次前20句子长度分布情况－总(以“词”为单位)

序号	句子长度(词)	频次(次)	占比(%)
1	15	2 834	2.98
2	18	2 708	2.85
3	16	2 681	2.82
4	22	2 667	2.80
5	20	2 647	2.78
6	17	2 629	2.76
7	19	2 603	2.74
8	21	2 584	2.72
9	13	2 507	2.64
10	14	2 486	2.61
11	23	2 463	2.59
12	24	2 402	2.52
13	12	2 267	2.38
14	25	2 263	2.38
15	26	2 208	2.32
16	27	2 159	2.27
17	11	2 119	2.23
18	28	2 115	2.22
19	10	1 975	2.08
20	29	1 961	2.06

与1998年相同的是,前20的句子长度的句子数在总句子数中的占比均超过2%。从频次排21的句子长度开始,各个长度的句子的数目占总数目的比重都低于2%,前20的句子长度又可以分为频次大于2 500词和频次小于2 500词两个区间,第一区间——1-9名占总句子数比重25.03%,第二区间——10-20名占总句子数比重25.52%。将前20的句子长度按照句子长度所属区间可以分为1-10、11-20、21-30三个区间,前20中长度在1-10词的有10,长度在11-20词的句子均包含在其中,长度在21-30词的有21、22、23、24、25、26、27、28、29,可以看出句子长度也主要集中在11-30词之间。本研究进一步统计了6类句子在长度上的分布情况,具体如表10所示:

表10　2018年人民日报语料句子长度区间分布情况-分(以“词”为单位)

序号	句号		分号		感叹号		冒号		省略号		问号	
	句子长度(词)	频次(次)	句子长度(词)	频次(次)	句子长度(词)	频次(次)	句子长度(词)	频次(次)	句子长度(词)	频次(次)	句子长度(词)	频次(次)
1	18	2 345	15	218	7	47	3	412	22	10	9	141
2	15	2 323	16	199	6	44	2	280	21	9	7	116
3	22	2 309	14	178	15	40	7	235	24	8	6	114
4	16	2 280	18	177	10	38	4	234	9	7	11	110
5	20	2 280	12	174	12	38	6	224	13	7	8	104
6	19	2 274	13	174	9	37	9	188	15	7	13	98
7	21	2 263	11	172	11	36	5	184	16	7	10	95
8	17	2 242	19	166	5	33	8	181	12	6	12	92
9	23	2 155	24	166	8	32	12	160	17	6	14	86
10	24	2 131	20	164	13	31	10	156	29	6	15	86
11	13	2 071	22	163	16	29	11	156	10	5	5	77
12	14	2 066	17	160	4	28	15	144	11	5	20	63
13	26	1 997	23	152	19	26	13	138	20	5	4	62
14	25	1 981	21	151	14	25	14	135	25	5	17	62
15	27	1 951	10	145	22	25	18	132	27	5	16	58
16	28	1 888	25	134	24	25	16	129	31	5	18	51
17	12	1 821	27	132	25	23	17	123	32	5	21	47
18	29	1 778	9	127	17	22	20	100	35	5	22	44
19	30	1 661	26	127	3	21	19	90	18	4	19	42
20	11	1 630	8	119	18	20	22	83	19	4	3	39
	31	1 630	28	104	/	/	/	/	23	4	/	/
	/	/	/	/	/	/	/	/	38	4	/	/

将句子按照句号、分号、感叹号、冒号、省略号和问号来划分,按照前10的句子长度分布来看:以句号结尾的句子长度主要集中在15-24词;分号结尾的句子长度主要集中在11-24词;感叹号结尾的句子长度主要集中在5-15词;冒号结尾的句子长度主要集中在2-12词;省略号结尾的句子长度主要集中在9-29词;问号结尾的句子长度主要集中在6-15词。

与整体的前20名高频句子长度分布相比较,句号结尾的句子长度分布与整体分布基本一致,以除句号外的其他符号结尾的句子的频次排在前20名的句子长度数与整体相差较大,但由于数目较少,所以对最终的分布情况影响较小。以分号、感叹号、冒号、省略号和问号这些符号结尾的句子虽然数目较少,但其高频句子长度仍在句子长度最集中的区间内。

以词为单位统计句子长度对1998年和2018年人民日报语料进行分析,本研究发现以下现象:比较表7和表9可以发现,2018年语料量虽然较1998年语料量有很大的提升,但是以词为单位统计句子长度,句子长度除极个别有所增长外,在整体长度分布上基本一致。1998年和2018年语料的句子主要长度分布在9-28词之间,最高频次出现在11-29词之间,整体呈现为先上升再下降的趋势。句子数量主要集中在一个长度

较小区间内,但是句子长度整体的跨度非常大,因此会有很长的"拖尾"现象,存在个别非常长的句子,但是这种句子占的比重非常小,基本上可以忽略不计。比较1998年与2018年人民日报语料各句子类别的句子量,除省略号外的类别句子数目均有增长,但省略号类别的句子数目有所减少。语料中最多的是句号结尾的句子,其他符号结尾的句子量较少,对整体分布情况的影响较小,整个语料的句子长度分布主要受句号结尾句子影响,基本与句号结尾句子的分布情况相同。以分号、感叹号、冒号、省略号和问号结尾的句子虽然较少,但其高频句子长度依然在整体高频句长区间内。

4 词分布上的齐普夫定律验证

齐普夫博士在对大量文本数据进行词频统计的研究中,提出以下词频分布规律:$f = Cr^{-\alpha}$,其中 f 是词频,r 表示词频的排序序号,C 和 λ 是参数,得到齐普夫表达式的一般表达形式 $F \times R = C$。

在公式 $f = Cr^{-\alpha}$ 中,如将 f 和 r 放在双对数坐标系中时,$\log(f) = \log(C) - \alpha\log(r)$ 所绘出的曲线接近一条直线,且斜率近似为 -1,即 α 的值接近1。后来的学者们在大量数据的基础上进行进一步研究,发现上述公式并不能完全地反映频率词典中词频的分布规律。如 r 的值与 f 的值之间存在唯一对应关系,这与现实情况中不同词拥有相同词频的现象不符,实验证明,当 $15 < r < 1500$ 的时候,频率相同的词群容量不大,当 $r > 1500$ 时,即单词的频率较小时,频率相同的词群的容量会陡增,引发数据稀疏问题。所以,齐普夫定律的适用情况仍具有探索和研究的空间。在上述对句子长度进行以字和词为单位的统计分析基础上,本研究结合齐普夫定律,进一步地从词的静态分布上对词的分布情况进行统计和分析。本文对1998年1月和2018年1月的人民日报语料分别进行词频统计,运用公式 $\log(f) = \log(C) - \alpha\log(r)$,借助 SPSS 工具验证齐普夫定律,在直角坐标系中绘图,α 即直线的斜率,$\log(C)$ 是拟合直线在 y 轴上的截距。具体流程如下:

1、对人民日报语料的频次和排序两列数据分别进行取对数处理;

2、借助于 SPSS 工具,使用线性回归分析,计算相关参数;

3、根据两列数据画出图形,绘制拟合直线。

在上述流程的基础上,本研究分别得到1998年1月和2018年1月人民日报语料中词汇的分布情况,如表11、图3和图4所示:

表11 采用最小二乘法(OLS)对1998年1月和2018年1月人民日报语料的齐普夫定律线性回归拟合结果

语料	α	log(C)	R^2
1998年1月人民日报	1.331	14.222	0.975
2018年1月人民日报	1.417	15.663	0.976

表11展示出了 log(r) 和 log(f) 的线性回归拟合结果,可以看到:1998年1月人民日报语料回归分析的 $R^2 = 0.975$,表示自变量对因变量的解释能力达到了97.5%。拟合回归方程:$y = -1.331x + 14.222$。如图3,其中 x 轴为排序的对数,y 轴为词频的对数值。

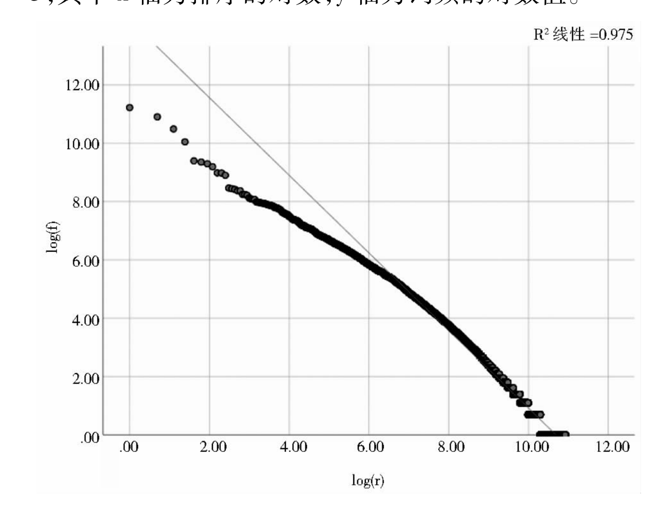


图3 1998年1月人民日报语料基于 OLS 的排序-词频分布以及拟合曲线

2018年1月人民日报语料回归分析的 $R^2 = 0.976$,表示自变量对因变量的解释能力达到了97.6%。拟合回归方程:$y = -1.417x + 15.663$,如图4所示,图中 x 轴为排序的对数,y 轴为词频的对数值。

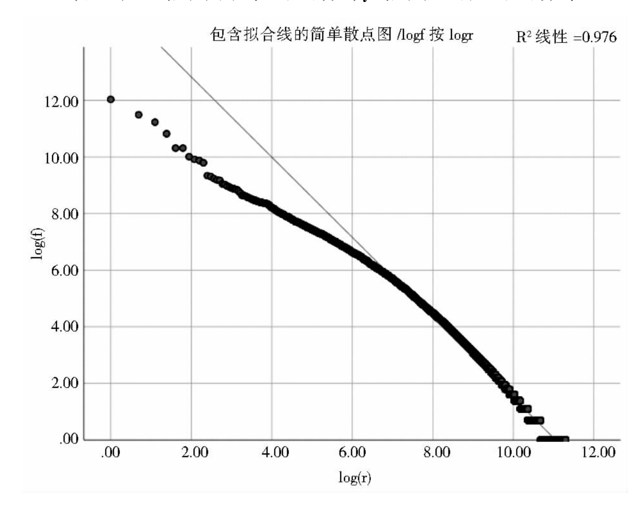


图4 2018年1月人民日报语料基于 OLS 的排序-词频分布以及拟合曲线

从图3和图4可以看出,1998年1月和2018年1月人民日报的两组语料在词频分布上表现十分相似,一个词出现的频率与它在频率表里的排名成反比,符合齐普夫定律。同时,从1998年到2018年,随着时代的变迁,人民日报的用词也出现了细微变化,两者拟合的曲线并非完全一致。斜率,即α值,是区分语言分布特征的重要参数,在大多数欧洲语言中,α取值为1,由于中英语言特征的差异,在1998年1月和2018年1月人民日报语料中α的取值分别为1.331和1.447,具有相对较大的差异,说明同一种语言在不同时期呈现出不同的词频分布特征,因此,简单地用1998年1月人民日报语料的语言分布规律和特征已经无法直接分析现在的人民日报语料。表12分别列举了1998年1月和2018年1月人民日报语料中相对词频最高的前20个词。

表12 1998年1月和2018年1月人民日报语料的词频分布结果

序号	1998年	2018年
1	的	的
2	在	在
3	了	和
4	和	了
5	是	是
6	一	中国
7	为	年
8	有	一
9	不	为
10	上	发展
11	对	不
12	中	新
13	中国	等
14	发展	有
15	这	对
16	要	中
17	多	上
18	与	也
19	到	与
20	他	要

从表12可以看出,相对词频排名前20的词基本重合,“的”“了”“在”等虚词和介词,都出现在两个语料词频排名前20的名单中,符合“最省力法则”。相较于1998年1月人民日报语料,2018年人民日报语料中“中国”“发展”“新”等词出现的相对词频明显上升,体现了人民日报在现代汉语的使用习惯上的变迁。

一般情况下,齐普夫定律较符合西方文献中词频分布的实际情况,但是,词频分布问题是很复杂的,使得齐普夫定律在适用范围上有一定的局限性,如图3和图4所示,数据在中段拟合效果最好,但前段和后段有不少数据偏离了拟合线,正如前文所述,尤其对出现频次特别高的词和特别低的词,并不能完全反映其词频分布特征。

5 结语

为了更深入和系统地研究人民日报语料在句子和词汇层级上的语言特征和风格,本文基于2018年1月人民日报分词语料,结合1998年1月人民日报分词语料对人民日报的句子长度和词频分布情况进行了研究。在统计和分析过程中,确定了人民日报语料的句子类型,并全面对比和分析了两个语料在句子长度和词频分布上的整体相同点和具体不同点。为了更加深入地统计和挖掘人民日报这一文本中所蕴含的语言规律和语言特征知识,在后续的研究中,一方面从词汇上深入地探究人名、地名、机构和时间等实体这一语言单位上的分布规律,另一方面融入句法的相应知识和技术,更加深入而细致地探究句子的分布特征。

参考文献:

[1] 黄水清,王东波. 新时代人民日报分词语料库构建、性能及应用(一)——语料库构建及测评[J]. 图书情报工作, 2019,63(22): 5-12.

[2] 黄水清,王东波. 新时代人民日报分词语料库构建、性能及应用(二)——句长与词的分析比较[J]. 图书情报工作, 2019,63(23):5-12.

[3] CLAYMAN D. Sentence length in Greek hexameter poetry[J]. Hexameter studies. quantitative linguistics, 1981(11): 107-136.

[4] 黄自然. 以“字”为单位的汉语平均句长与句长分布研究[J]. 齐齐哈尔大学学报(哲学社会科学版), 2018(1): 133-138.

[5] 李青苗. 从《左传》的偏正结构和句子长度看现代汉语细节意义的增强[J]. 东北师大学报(哲学社会科学版), 2018(4): 99-103.

[6] 王萍, 石锋. 汉语普通话不同语句类型的时长分布模式[J]. 语言教学与研究, 2019(2): 101-112.

[7] 左思民. 汉语句长的制约因素[J]. 汉语学习, 1992(3): 16-21.

[8] 张绍麒, 李明. 小说与政论文言语风格异同的计算机统计(实验报告)[J]. 天津师范大学学报: 社会科学版, 1986(4): 82-86.

[9] 黄自然, 贾成南. 平均句长在语言习得研究中的应用与问题[J]. 长江大学学报: 社会科学版, 2013(1): 95-97.

[10] 金志娟, 金星明. 学龄前儿童普通话平均句子长度和词汇广度

研究[J]. 中国循证儿科杂志, 2008, 3(4): 261－266.
[11] MONTGOMERY M, MONTGOMERY A, STEPHENS M. Sentence repetition in preschoolers: effects of length, complexity, and word familiarity[J]. Journal of psycholinguistic research, 1978, 7(6): 435－452.
[12] 李建平，张晓蕗. 中美中学生英语写作句子长度对比分析——一项基于高考英语作文的研究[J]. 教育测量与评价: 理论版, 2015 (7): 50－53.
[13] ZIPF G K. Human behaviour and the principle of least-effort [M]. Cambridge: Addison-Wesley, 1949.
[14] 冯志伟. 齐普夫定律的来龙去脉[J]. 情报科学, 1983 (2): 37－42.
[15] 孙清兰，王肇建. 齐夫定律的词等级确定方法探讨[J]. 东北师大学报: 自然科学版, 1993 (3): 32－37.
[16] 孙清兰. 高频，低频词的界分及词频估计方法[J]. 情报科学, 1992, 13(2): 28－32.
[17] LI W. Zipf's law everywhere[J]. Glottometrics, 2002(5): 14－21.
[18] TUZZI A, POPESCUI I, ALATMANN G. Zipf's laws in Italian texts[J]. Journal of quantitative linguistics, 2009, 16(4): 354－367.
[19] 沈关龙. 齐普夫定律与专题文献标题词频的研究及应用[J]. 情报理论与实践, 1988(2):58－64,130.
[20] 王崇德，来玲. 汉语文集的齐夫分布[J]. 情报科学, 1989, 10(2): 1－8.
[21] 何凤远. 中文词频分布与齐夫定律的汉语适用性初探[J]. 现代语文(语言研究), 2010(10):110－111.
[22] GABAIX X. Zipf's law for cities: an explanation[J]. The quarterly journal of economics, 1999, 114(3): 739－767.
[23] AXTELL R L. Zipf distribution of US firm sizes[J]. Science, 2001, 293(5536): 1818－1820.
[24] ADAMIC L A, Huberman B A. Zipf's law and the Internet[J]. Glottometrics, 2002, 3(1): 143－150.

作者贡献说明：

黄水清:提出相关概念及整体研究思路,修订完稿;
王东波:数据处理及初稿撰写。

Construction, Performance and Application of New Era People's Daily Segmented Corpus (Ⅲ)
——Analysis and Comparison of Sentence Length and Word

Huang Shuiqing[1,2]　Wang Dongbo[1,2]
[1] College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095
[2] Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] The statistics and analysis of sentence length in different dimensions and vocabulary distribution based on the New Era People's Daily(NEPD) word segmentation corpus is not only conducive to a relatively comprehensively and systematically understanding of the linguistic characteristics of the contemporary Chinese text, but also beneficial to the subsequent exploration of natural language processing and text mining of the text. [**Method/process**] Based on the word segmentation data of People's Daily in January 2018 and the word segmentation data of People's Daily in January 1998, 6 sentence categories used in the statistics were determined, and the sentence length distribution of character and word units was counted and analyzed, and the distribution of words in static state was revealed based on Zipf's law. [**Result/conclusion**] From the perspective of the sentence length distribution in the word dimension and the Zipf distribution of vocabulary, the sentence length and vocabulary distribution have both changed in the 1998 and 2018 corpora as time goes by, but this change is continuous and related.

Keywords: New Era People's Daily segmented corpus　segmented corpus　sentence length　distribution of word　Zipf's law